POLITECNICO
MILANO 1863

COVID 19 - Uno studio del Politecnico di Milano per capire i segreti delle sequenze virali

Il motore di ricerca ViruSurf svela come cambia il genoma del virus responsabile della pandemia

Milano, 23 novembre 2020 – Dall'inizio del 2020, i laboratori di tutto il mondo sequenziano materiale genetico che deriva dai tamponi positivi di persone affette da COVID-19 e depositano poi le sequenze virali in tre principali banche dati: GenBank, COG-UK e GISAID. Per muoversi agilmente in questa enorme mole di dati e "surfare" alla ricerca di connessioni utili alla comprensione del virus, il gruppo di ricerca del Politecnico di Milano guidato dal Prof. Stefano Ceri ha realizzato **ViruSurf** (http://gmql.eu/virusurf), un motore di ricerca che si avvale di un database centralizzato collocato al Politecnico. Il database viene aggiornato periodicamente e ad oggi contiene 200,516 sequenze di SARS-CoV-2, il virus responsabile della pandemia, e 33,256 sequenze di altre specie, anch'esse associate ad epidemie di interesse per l'uomo, tra cui SARS, MERS, Ebola e Dengue.
Ogni sequenza è descritta secondo quattro prospettive: le caratteristiche del virus e dell'organismo ospite, la tecnologia utilizzata, il progetto di sequenziamento, le mutazioni dei nucleotidi e degli amino acidi che si trovano in diversi geni. Il vantaggio di ViruSurf è di includere un algoritmo che calcola le mutazioni virali in maniera omogenea, ovvero indipendente dalla loro provenienza, gestito su cloud per ridurre i tempi di esecuzione. Il database è ottimizzato per offrire risposte istantanee agli utilizzatori del motore di ricerca.

Tra i diversi sviluppi futuri di ViruSurf, il più importante, finanziato da **EIT Digital** con un progetto semestrale, è un servizio informatico per elaborare nuove sequenze virali identificando in esse particolari mutazioni associate a maggiore o minore severità e virulenza. Utilizzato in campo medico, in fasi meno acute della pandemia, permetterà di arricchire la "cartella clinica" del paziente con la sequenza del virus che lo ha infettato. Sarà inoltre possibile utilizzare ViruSurf per il monitoraggio dei virus nella gestione di allevamenti e coltivazioni. Il sistema consentirà a breve di tracciare gli epitopi – sequenze di amino acidi del virus che sono critiche per lo sviluppo di vaccini – ad esempio per trovare, per ogni epitopo, le mutazioni della sua sequenza diffuse in alcune regioni del pianeta, che potrebbero pregiudicare l'efficacia del vaccino.

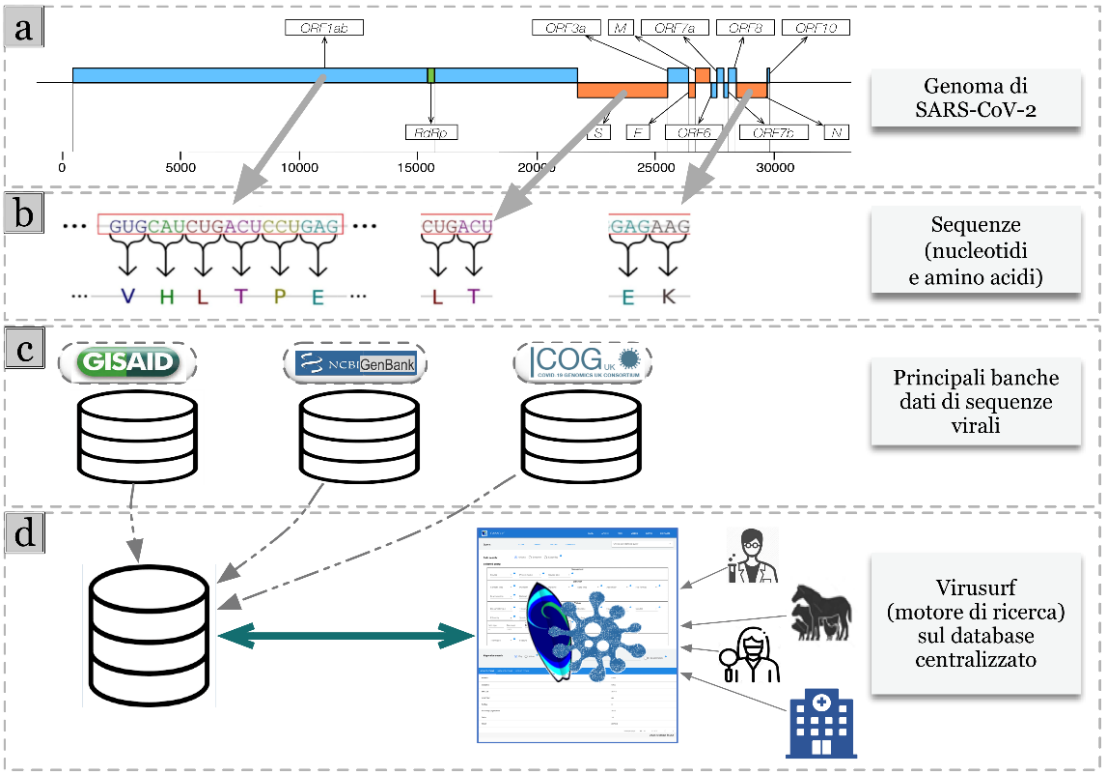
Ufficio Relazioni con i Media
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano

T +39 02 2399 2441
C. +39 3666211435
relazionimedia@polimi.it
www.polimi.it

POLITECNICO
MILANO 1863

“Nel progetto GeCo, finanziato da **European Research Council**, avevamo già sviluppato un motore di ricerca per il genoma umano, chiamato GenoSurf; ad inizio pandemia non esisteva un analogo sistema per le sequenze virali. Per comprenderne i requisiti, abbiamo intervistato venti esperti virologi da tutto il mondo. Il risultato è un sistema di semplice utilizzo: chiunque può collegarsi e capire, ad esempio, quando una mutazione virale è apparsa per la prima volta e come si è diffusa nel mondo”—racconta **Stefano Ceri**, leader del progetto. L’articolo è pubblicato su una rivista di grande rilievo, Nucleic Acids Research (https://doi.org/10.1093/nar/gkaa846), che raccoglie annualmente i database più importanti per la biologia. Hanno contribuito all’articolo anche **Pietro Pinoli**, progettista degli algoritmi, **Arif Canakoglu**, software architect, **Anna Bernasconi**, data designer, **Tommaso Alfonsi,** responsabile della acquisizione dei dati, e **Damianos P. Melidis** di L3S (Hannover), autore di alcuni algoritmi.

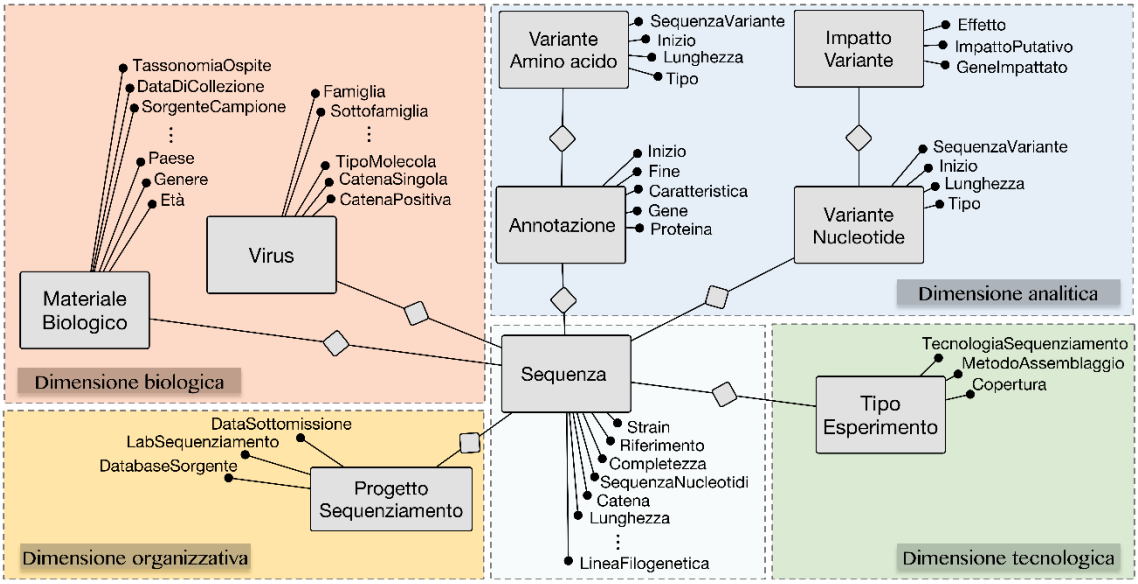
Link alla video-presentazione di Anna Bernasconi al Congresso ER2020 (6 Novembre 2020) https://youtu.be/HjnEOQnUnEg



Ufficio Relazioni con i Media
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano

T +39 02 2399 2441
C. +39 3666211435
relazionimedia@polimi.it
www.polimi.it

Dal genoma del virus SARS-CoV-2 (a) si estrae la sua sequenza di nucleotidi e amino acidi (b); le sequenze, depositate nelle banche dati mondiali: GENBANK, GISAID, COG-UK (c), sono importate nel database centralizzato del Politecnico, su cui opera Il motore di ricerca ViruSurf (d).



Schema del database integrato: le sequenze del virus vengono descritte in base alle loro caratteristiche biologiche (specie virale e ospite), al progetto che le ha prodotte, alla tecnologia di sequenziamento e alle proprietà del genoma (annotazioni, mutazioni della sequenza dei nucleotidi e degli amino acidi).

ENGLISH VERSION

COVID 19 – A research of Politecnico di Milano for discovering the secrets of viral sequences

The search engine **ViruSurf** discloses the changes of the genome of the virus responsible for the pandemics

Milan, 23 november 2020 – Since the beginning of 2020, labs from all around the world are sequencing the material from positive tests of people affected by COVID-19 and then depositing sequences mostly to three points of collection: GenBank, COG-UK, and GISAID. A fast exploration of this huge amount of data is important for understanding how the genome of the virus is changing. For enabling fast “surfing” over this data, the research group of Politecnico di Milano led by Prof. Stefano Ceri has developed **ViruSurf** (http://gmql.eu/virusurf), a search engine operating on top of a centralized database stored at Politecnico. The database is

Ufficio Relazioni con i Media
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano

T +39 02 2399 2441
C. +39 3666211435
relazionimedia@polimi.it
www.polimi.it

POLITECNICO
MILANO 1863

periodically reloaded from the three sources and as of today contains 200,516 sequences of SARS-CoV-2, the virus causing COVID-19, and 33,256 sequences of other viral species also associated to epidemics affecting humans, such as SARS, MERS, Ebola, and Dengue.

Every sequence is described from four perspectives: the biological features of the virus and the host, the sequencing technology, the project that has produced the original data, the mutations of the whole sequence of nucleotides and of gene-specific amino acids. The advantage provided by ViruSurf is the use of an algorithm for computing viral mutations homogeneously across sources, using cloud computing. The database is optimized for giving quick responses to the search engine surfers.

Among the future developments of ViruSurf, the most important, funded by a six-month-long project by **EIT Digital**, is a bio-informatic service for ingesting new viral sequences, which highlights the presence of viral mutations associated with enhanced or reduced severity and virulence as they are discovered. Used in clinics, particularly with a less acute pandemics spreading, it will support the addition of critical information to the patient health record; other uses will be possible in the context of animal farming or of the food chain. The system will soon allow the tracing of epitopes – amino acid sequences that are used in vaccine design – for instance to associate epitopes with mutations of the virus that could be present in given countries of the world and that could affect vaccine.

"In the GeCo project, financed by the **European Research Council**, we had already developed a search engine for datasets describing the human genome, called GenoSurf; at the beginning of the pandemic, there was no such system for viral sequences. To better understand its requirements, we interviewed about twenty expert virologists from all over the world. The result is a user-friendly system: any researcher can connect to it and perform queries, for instance, about when a viral mutation started and how it has spread in the world"—says **Stefano Ceri**, the project leader. The article is published on a high relevance journal, Nucleic Acids Research (https://doi.org/10.1093/nar/gkaa846), in the database issue that every year collects the descriptions of the most significant biological databases. The article is authored also by **Pietro Pinoli**, algorithm designer, **Arif Canakoglu**, software architect, **Anna Bernasconi**, data designer, **Tommaso Alfonsi,** designer of the data loading pipeline, and **Damianos P. Melidis** from L3S (Hannover), author of some algorithms.

Link to the video of Anna Bernasconi's presentation at the ER2020 Conference ER2020 (Nov. 6, 2020) https://youtu.be/HjnEOQnUnEg

Ufficio Relazioni con i Media
Politecnico di Milano
Piazza Leonardo da Vinci 32
20133 Milano

T +39 02 2399 2441
C. +39 3666211435
relazionimedia@polimi.it
www.polimi.it